Analysis of High Throughput Flow Cytometry Data using *plateCore*

March 31, 2009

Abstract

Flow cytometry (FCM) software packages from R/Bioconductor, such as flowCore and flowViz, serve as an open platform for development of new analysis tools and methods. We created *plateCore*, a new package that extends the functionality in these core packages to enable automated isotype-based gating and make the processing and analysis of plate-based data sets from high throughput FCM screening experiments easier. *plateCore* was used to analyze data from a BD FACS CAP screening experiment where 5 Peripheral Blood Mononucleocyte Cell (PBMC) samples were assayed for 189 different human cell surface markers. This same dataset was also manually analyzed by a cytometry expert using the FlowJo data analysis software package (TreeStar, Ashland OR). We show that the expression values for markers characterized using *plateCore* are in good agreement with those from the manual analysis.

Introduction

While there are a number of different software packages available for analysis of FCM data, these programs are often ill-suited to the development of new methods needed for analyzing high-throughput FCM studies. Flow Cytometry High Content Screening (FC-HCS) experiments generate large volumes of data, which requires a systematic approach to preprocessing, gating (i.e. filtering), and summarizing results for robust analyses. Automation of these steps would allow analysis pipelines to be robust, objective, and match the high-throughput capacity of modern cytometers. Unfortunately, current approaches to FC-HCS analysis are semi-automated at best, and they often require significant subjective and error-prone manual intervention to identify cells of interest (Maecker et al., 2005). It is therefore desirable to develop programmatic approaches to process FCM data.

FCM packages available through the Bioconductor (Gentleman et al., 2004) project provide an open platform that can be used by cytometrists, bioinformaticians, and statisticians to collaboratively develop new methods for automated FC-HCS analysis. The basic data processing tools for importing, transforming, gating, and organizing raw FCM data are in the *flowCore* package (Hahne et al., 2009), and the visualization functions are in *flowViz* (Sarkar et al., 2008). The Bioconductor model for FCM data analysis facilitates the development of new analysis methods, since the overhead associated with accessing and visualizing FCM data is handled by *flowCore* and *flowViz*. The availability of *flowCore* and *flowViz* has enabled the creation of new tools for quality assessment of large FCM experiments, such as *flowQ* (Gentleman et al.), and model-based clustering and automated gating, such as *flowClust* (Lo et al., 2008).

We have developed an R package (*plateCore*) that also takes advantage of the functionality in *flowCore* and *flowViz* to create methods and data structures for processing large, plate-based FCM datasets. Additionally, we have implemented new tools to make it easier to integrate textual descriptions of plate layouts and also functions for automated gating based on non-parametric analysis of negative control wells. This study presents results from an automated *plateCore* analysis of a PBMC lymphocyte FACS CAP data set, which included 189 different antibody-dye conjugates and their controls arranged on a 96-well plate. The *plateCore* output was compared to an analysis by an expert cytometrist using FlowJo to evaluate the performance of the automated approach.

plateCore is not designed to be a GUI driven end-user tool, but rather to help develop a standardized platform for the analysis of FC-HCS data. These analyses often represent a collaborative effort between cytometry experts who generate the data and the quantitative individuals who help deal with the large volume information. In order for this collaboration to work, the cytometrists must have confidence in the results of the automated analysis. To this point, we demonstrate the equality of our results to those produced by an expert cytometrist using FlowJo.

Materials and Methods

Flow Cytometry Data

The data analyzed in this study was part of the initial set of experiments used to validate the BD FACS CAP platform. BD FACS CAP was designed as a cell characterization tool to screen for the presence of a large number of different human cell surface markers, and it was important to show that the assay was able to correctly identify positive and negatively staining markers on a well studied cell population, such as PBMC lymphocytes. Previously frozen PBMC samples from two donors were analyzed on a BD FACS Calibur using BD FACS CAP staining plates. The analysis was performed on 96-well plates with 189 different antibodies arrayed three per well in 63 test wells, along with 30 isotype control wells and three unstained controls. The complete list of BD FACS CAP antibodies can be found at http://www.bd.com/technologies/discovery_platform/BD_FACS_CAP.asp. FCM files for the 5 plates (two for Donor 1 and three for Donor 2), are available for download from http://www.ficcs.org.

plateCore

The *plateCore* scripts used to perform the analysis are provided in supplementary materials. Briefly, the FCM files were first processed using a combination of static gates (`rectangleGate`) and data driven gates (using `norm2filter` and *flowCore*) to pick out the lymphocytes in the forward (FSC) and side scatter (SSC) channels. The quality of the data was then assessed by looking for fluidic events such as bubbles, pressure drops, or large aggregates that can shift the baseline fluorescence readings. Fluidic events can often be identified by plotting the empirical cumulative density (ecdf) plots of FSC values for each well, and looking for distributions shifted relative to other wells (Le Meur et al., 2007). Based on the ecdf plots, several wells were further investigated by cytometry experts who determined that the shifts were in an acceptable range. Next the threshold between positive and negative cells were determined using the isoytpe controls, which provided a gross estimate of non-specific binding in the primary antibodies. One-dimensional gates were created using using the isotype thresholds, and these gates were applied to identify cells that are positively stained for each marker.

An example of the progression from raw FCM data files to a completed *plateCore* analysis is shown in Figure 1. List mode FCS files for a single plate were read into a `flowSet` using *flowCore*, and then a `flowPlate` was created by integrating the plate annotation file with the `flowSet`. The `flowPlate` was then compensated, data quality was assessed, and gates were set according to a negative control. These control gates were then applied to test wells to find cells that had specific staining in channels of interest.

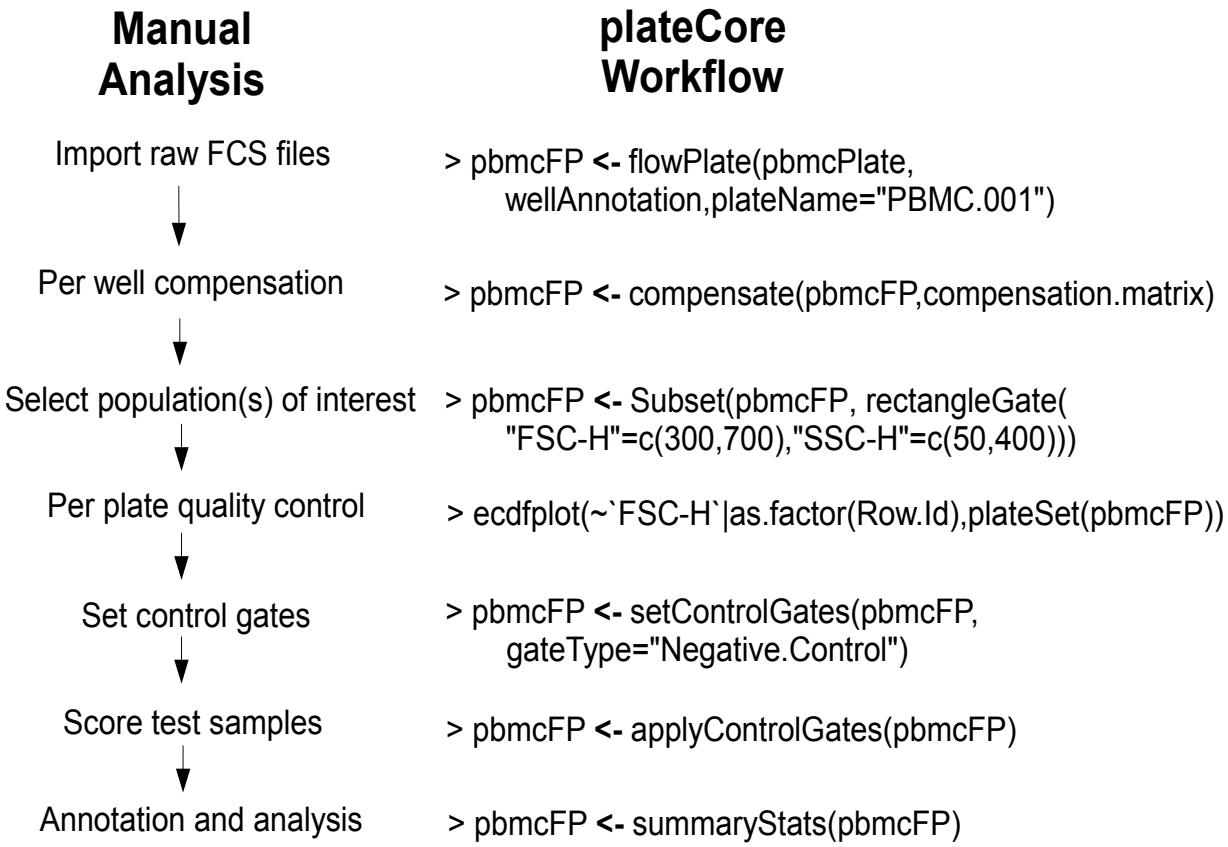

Figure 1: Typical FC-HCS plate workflow on the left and corresponding steps from a PBMC lymphocyte *plateCore* analysis on the right. Compared to analyses performed using existing GUI FCM tools, *plateCore* can reduce the level of subjectivity associated with creating the negative control gates and also makes it easier to aggregate multiple plates into an experiment level object for visualization and reporting. Providing *plateCore* scripts along with the raw FCM data for FC-HCS experiments helps ensure that an analysis is transparent and reproducible.

FlowJo

In addition to *plateCore*, the five PBMC plates were also analyzed using FlowJo. First, an analysis template was created where test wells and their corresponding isotype control well were assigned to one of 30 groups. Wells in each group had similar sets of antibody-dye conjugates, and the expression threshold (*i.e.*, isotype gate) was initially set using the isotype control well. Data for each plate was imported into FlowJo using the template and lymphocytes were selected using a morphology (FSC-SSC) gate. Event data for the isotype well was then visualized on a log scale, and the expression threshold for each stained channel was set by picking a value that lies above the bulk of the events. For BD FACS CAP, the isotype gates were initially set so that approximately 1% or less of the events in the isotype well were above the threshold. These gates were then applied to the test wells, and the gates were moved up or down depending upon positive and negative test well populations. If the the population of cells in positive wells was much higher than the isotype gate, then the gate was moved up to help reduce false positives associated with non-specific staining. Similarly, if the isotype gate was higher than negative samples, the gate would be moved down to ensure that positive cells were classified correctly. The percentage of cells above the threshold for each of the 189 antibodies was then exported for each plate.

Results

Although this study focuses on comparing two different FC-HCS analysis methods, it is important to consider the original goal of the experiment used to generate the data when interpreting the results. BD FACS CAP was designed to provide a standard assay platform for screening a large number of markers on many different cell types. The validation effort for BD FACS CAP included running the assay on well-charaterized cell types to find markers with either positive or negative staining, and comparing these results to published cell expression profiles in the literature. The PBMC lymphocyte staining results presented in the following section represent one of the cell types used for validating the technology.

FlowJo Output

Descriptions of marker expression profiles for particular cell populations in flow cytometry often use terms like positive-negative, or bright-dim, to qualify the amount of target present. Since FACS CAP is a standard platform for screening a wide range of cell types, and antibody concentrations were not optimized for these PMBC samples, results are reported as the percentage of cells above the isotype gate rather than positive or negative. Follow-up studies, including single color titrations and competition experiments, are needed to definitively show that a marker is present. These additional analyses of markers that have been characterized using FACS CAP show that markers with $\geq 90\%$

of the cells above the isotype threshold are usually confirmed as positive, while staining in markers with $\leq 10\%$ of cells above the isotype threshold is often the result of non-specific binding (data not shown). Note that these percentages refer to the fraction of cells above the isotype threshold, but this does not necessarily imply heterogeneous staining in multiple populations.

Automating the creation and modification of isotype gates made by cytometrists analyzing BD FACS CAP data using FlowJo is challenging. Cytometrists adjust gates based on expert knowledge about the performance of specific antibody types and dyes, or after identifying positive or negative test samples. If the isotype gate cut off the bottom portion of a positive cell population in a test well, then the gate was moved down. Similarly, if the the isotype gate included too many cells from negative test wells, it was moved up. Results from the FlowJo based gating of replicate PBMC plates are shown in Figure 2. Detailed results for each marker are not presented in this study, but since the majority of antibodies on the FACS CAP staining plate are known to bind different leukocytes, it is not surprising that a large fraction would be identified as positive on PBMCs. Markers such as CD44, CD45, CD47, and CD59 are broadly expressed on lymphocytes and were positive (>99%) in this study.

plateCore versus FlowJo

The automated approach to gating in *plateCore* determines the threshold using isotype control wells. The gate (G_{ij}) for isotype i, channel j is set according to:

$$G_{ij} = MFI_{ij} + 4MAD_{ij}, \tag{1}$$

where MFI is the Median Fluoresence Intensity and MAD is Median Absolute Deviations on a linear scale. The choice of 4 MADS is an attempt to set the gate above the 99th percentile of cells in the isotype stained wells and replicate the actions of the cytometrists when initially creating the isotype gates. Empirical evidence from analyses of additional FACS CAP experiments not given here show that the 4 MADS setting produces gates that are very similar to those made by expert cytometrists when analyzing PBMC cells. While this simple, non-parametric method works surprisingly well for BD FACS CAP, advances in model-based clustering methods, such as those in *flowClust*, should lead to future performance improvements in automated gating.

Comparisons of the output from the *plateCore* and FlowJo analyses are shown in Figure 3. Both methods produce nearly identical estimates for markers that were either clearly positive ($\geq$99%) or clearly negative ($\leq$1%). These cell populations are not close to the isotype threshold, and therefore different isotype gate settings have little or no effect on estimates of the percentage of cells above the gate. In situations where the isotype gate splits a test cell population, small changes to the gate can dramatically change these estimates. This effect is evident in the results from replicate plates using FlowJo (Figure 2), and also in comparisons of FlowJo and plateCore (Figure 3), where

estimates for markers having approximately 50% of the cells above the isotype gate are more variable than markers having $\leq$1% or $\geq$99%.

Looking in detail at one marker, CD112 (Figure 3), where FlowJo and *plateCore* gave very different answers we see an example of a gate that was updated in the manual analysis based on a negative sample. Figure 4 shows the *plateCore* and FlowJo isotype gates for CD112 (IgG1-PE) and CD109 (IgG1-PE). The *plateCore* gate was set using the isotype signals, while the FlowJo gate was moved upwards based on staining of CD109.

Gating Quality Assessment

Since we may not always have access to output from expert cytometrists to help determine if our automated gating is reasonable, we need an alternative approach to assessing the quality of our isotype-based gates. The strategy we implemented in *plateCore* looks at the percentage of cells above the isotype gates versus the Median Fluorsence Intensity (MFI) ratio to see if the gating was consistent across the experiment. The MFI ratio is defined as the ratio of the median fluoresence signal for a marker divided by the median signal for its isotype control. Essentially, the MFI ratio tells us how well separated the stained test sample signal is from its negative isotype control, and this separation is clearly related to the percentage of cells above the isotype gate (Figure 5).

Figure 5 shows the results of a robust logistic regression for the percentage of positive cells on the MFI ratio for the 189 markers from the 5 plates. The bulk of the marker values (927 out of 945) are within 2 standard residuals from the best fit line, which is surprising since the 189 different antibodies were conjugated to different fluorophores (either Alexa 488, FITC, PE, PerCP, APC, or Alexa 647) and matched against different isotypes (either IgG1, IgG2, IgG2a, IgG2b, IgG3, or IgM). We expected that differences in fluorescence intensity between dyes, and variation in non-specific binding by different antibody types, would make direct comparisons difficult.

The 18 values that were more than 2 standard residuals away from the line were examined in detail and the isotype gate settings were found to be reasonable. In this case the flagging was the result of a positive and negative staining population of cells, which made the relationship between the MFI ratio and the fraction of cells above the isotype gate look very different than markers staining a single population. Detailed plots for one of the markers that was flagged, CD3, are shown in Figure 6.

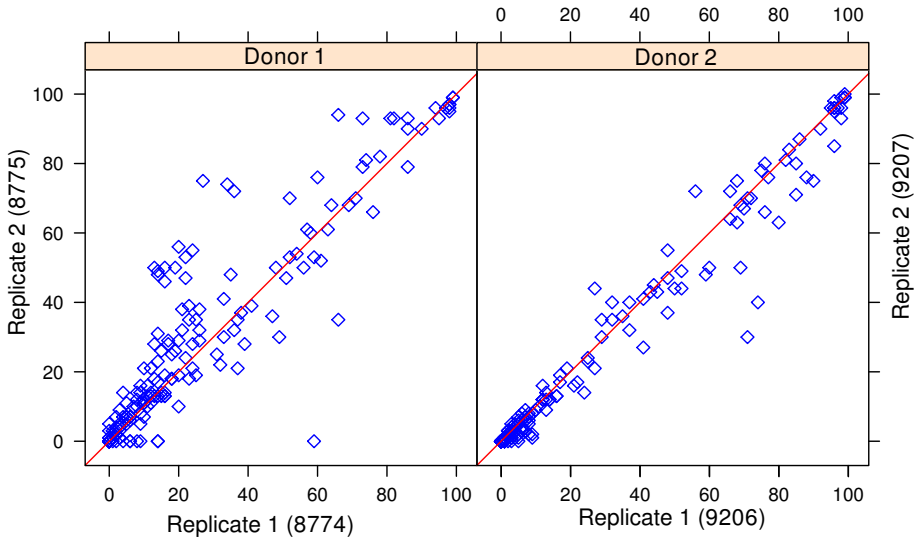

Figure 2: FlowJo estimates for the percentage of cells above the isotype threshold for 189 markers on replicate plates for donor 1 and donor 2. Estimates from markers where the center of the cell population was near the isotype threshold, around 50%, were more variable than samples which were clearly positive($\geq$99%) or negative ($\leq$1%). The correlation for replicate plates was strong in both donors, with donor 1 at 0.92 and donor 2 at 0.98. Plate 9208 for donor 2 is not shown, since the results are very similar to 9206 and 9207.

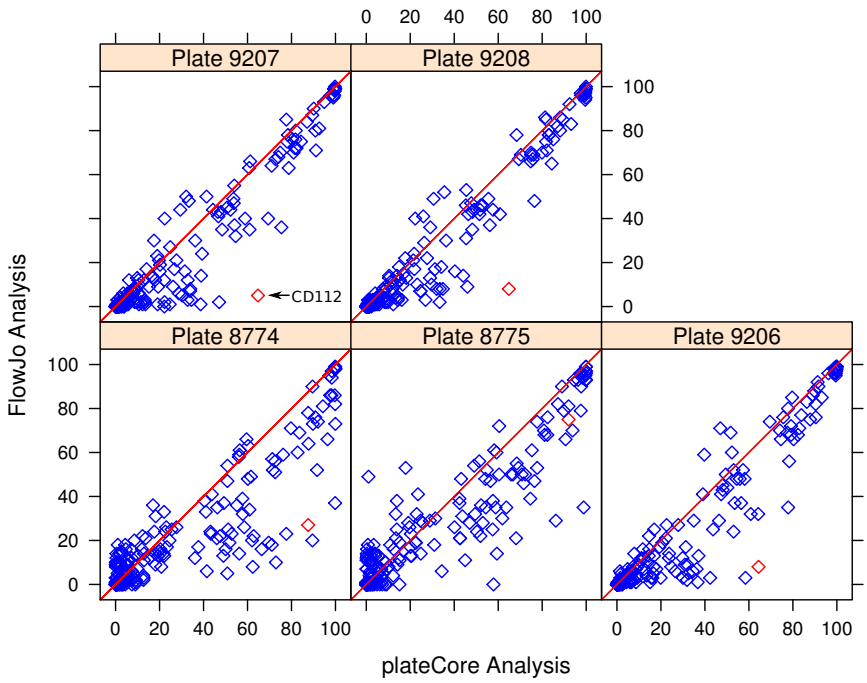

Figure 3: Plot showing the percentage of cells above the isotype threshold from plateCore (x-axis) and FlowJo (y-axis) for each of the 189 markers on the 5 PBMC plates. If the two methods produce similar estimates, then the values should be near the red line (y=x). In plateCore the isotype threshold was determined using only information from the isotype control well, while the threshold in FlowJo may be adjusted after identifying either positively or negatively staining test samples. Generally, these FlowJo adjustments resulted in the isotype gate being set a higher level to exclude a negative test sample. The effect of increasing the isotype threshold can be seen in these plots, where most disagreements are cases where plateCore estimates are higher than FlowJo. Detailed plots for one marker, CD112 (red diamond), where the two methods give different results are shown in Figure 4.

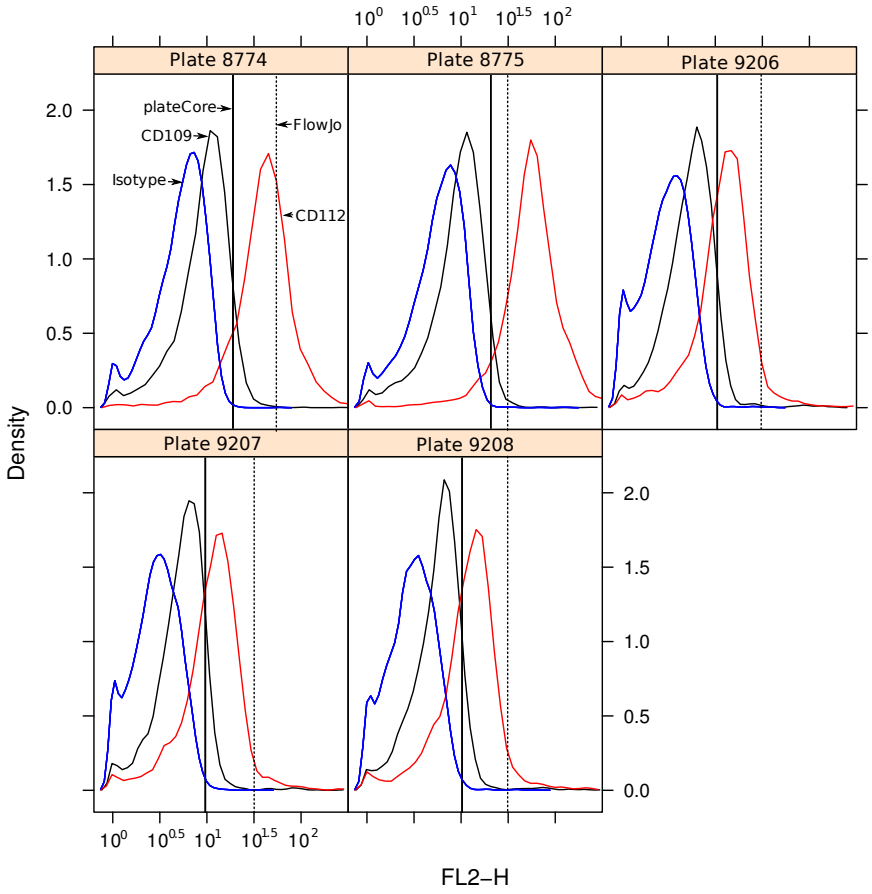

Figure 4: Density plots showing the plateCore (solid black) and FlowJo (dashed black) isotype gates for CD112 and CD109, which shared the same isotype control (IgG1-PE). The plateCore and FlowJo analyses gave different estimates for CD112 (see Figure 3), which was caused by the gate being moved higher in FlowJo based on the presumed negative staining for CD109.

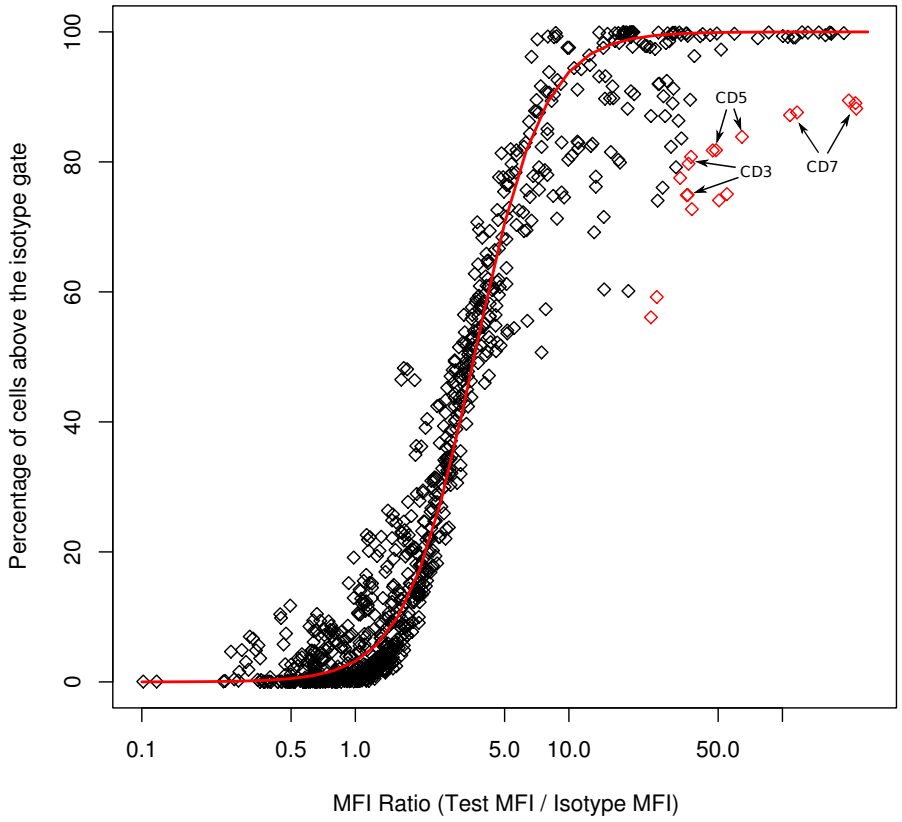

Figure 5: Quality of the automated gating was assessed by performing a robust logistic regression of the percentage of cells above the isotype gate on the log transformed MFI ratio, and looking for estimates that were more than 2 standardized residuals away from the best fit line (red line). There were 18 estimates flagged in this study (red diamonds) where the value was different than we would predict from the MFI ratio. Detailed examination of these 18 cases showed that the isotype gate settings were reasonable, but they differed from other markers in that they had more than one population of stained cells. Sample density plots for one of these markers, CD3, are provided in Figure 6.

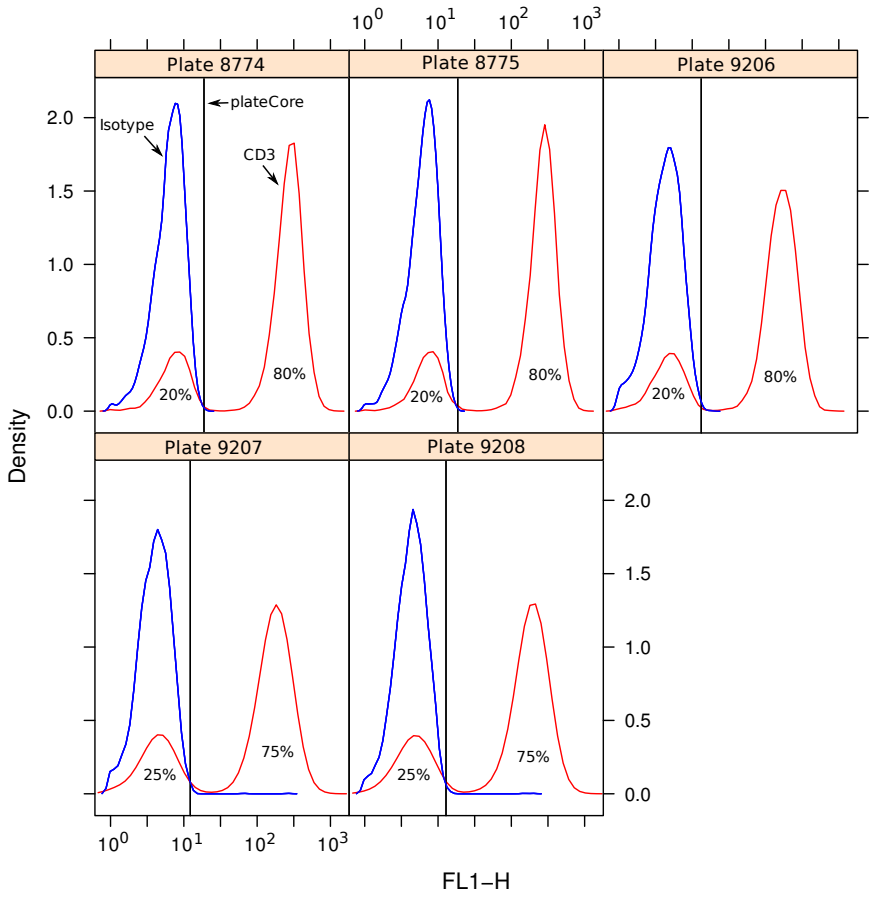

Figure 6: Density plot for CD3 (IgG1-Alexa 488), which was flagged for further evaluation by our gating quality assessement (Figure 5). The isotype gate setings look reasonable, however the MFI ratio for CD3 was very different from other markers that also had 75-80% of their cells above the isotype gate. Looking at Figure 5, other markers with 75-80% had MFI ratios near 5, while CD3 has an MFI ratio of 31-37. The flagging was the result of 2 cell populations for CD3, whereas most other markers stain a single population.

Discussion

We showed that the non-parametric approach to gating implemented in plateCore produces results that are similar to those from an expert user when analyzing PBMC lymphocyte FACS CAP data. Since this was a screening assay, the goal was to quickly and reproducibly process a large volume of data to get an approximate expression value for each of the 189 human cell surface markers, and then perform more in-depth analysis for markers that were of biological interest. Using *plateCore*, we were able to reduce the level subjectivity in setting isotype gates, eliminate mistakes associated with manual data annotation and export, and automate the creation of plots and data quality reports that summarized the experiment. Additionally, the *plateCore* scripts and experimental annotation can be shared with other cytometry groups, allowing them to reproduce our analysis.

Looking at markers where FlowJo and plateCore gave different results, such as CD112, it is not clear that either method gated the cells correctly. The gene for CD112 (PVRL2) has been shown to be expressed on a subset of B cells, CD4 T cells, CD8 T cells, and NK cells in healthy donors using microarrays (Critchley-Thorne, 2007), so the plateCore results showing 65-92% of the lymphocytes above the isotype gate may actually represent specific staining. Unfortunately, increasing the isotype (IGg1-PE) threshold in FlowJo to eliminate what looks like negative, non-specific staining in CD109 (Figure 4) also seems reasonable. More focused studies will have to be performed to determine if the staining for CD112, and other markers that disagreed, was positive or negative.

Another advantage of performing the analysis in plateCore is the ability to easily verify that isotype-based gate settings were consistent across the 5 PBMC plates. Figure 5 shows that there is a linear relationship between the distance between the test and control cell populations, as measured by the MFI ratio, and the fraction of cells above the isotype gate. Deviation from this line can indicate either a problem with the isotype gate or that the sample has multiple cell populations (Figure 6). We note that this approach does not actually tell us if the gating is correct, simply whether or not the isotype gating was consistent.

The complexity of large FCM experiments, like BD FACS CAP, highlight the difficulty of applying existing FCM analysis platforms to high-throughput studies. Generating and interpreting results from this PBMC study required extensive collaboration between flow cytometrists, bioinformaticians, and statisticians. At various points in the analysis, each group needed to access the raw data, annotation, and details about the experimental design. Providing this access using stand-alone FCM platforms is expensive in terms of the price of multiple software licenses and in time spent training statisticians and bioinformaticians to use the programs. Fortunately the Bioconductor FCM packages are modeled on standard data structures used for microarrays, which should already be familiar to most quantitative individuals working on high-throughput biological problems. We found that *flowCore*, *flowViz*, and *plateCore* provided an open analysis platform that facilitated communication between the flow cytometrists generating the

data and the computational experts analyzing the data.

References

N. Nacu S. Weber J. Holmes SP. Lee PP. Critchley-Thorne, RJ. Yan. Down-regulation of the interferon signaling pathway in t lymphocytes from patients with metastatic melanoma. *PLoS Med*, 4:e176, 2007.

R. Gentleman, F. Hahne, J. Kettman, and N. Le Meur. Bioconductor package flowq. URL `http://www.bioconductor.org`.

Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang, and Jianhua Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL `http://genomebiology.com/2004/5/10/R80`.

F. Hahne, N. LeMeur, R.R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, and R. Gentleman. flowcore a bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, 2009.

N. Le Meur, A. Rossini, M Gasparetto, C Smith, R.R. Brinkman, and R. Gentleman. Quality assessment of ungated flow cytometry data in high throughput experiments. *Cytometry A*, 71:393–403, 2007.

K. Lo, R. R. Brinkman, and R. Gottardo. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A*, 73A:321–332, 2008.

Holden T Maecker, Aline Rinfret, Patricia D'Souza, Janice Darden, Eva Roig, Claire Landry, Peter Hayes, Josephine Birungi, Omu Anzala, Miguel Garcia, Alexandre Harari, Ian Frank, Ruth Baydo, Megan Baker, Jennifer Holbrook, Janet Ottinger, Laurie Lamoreaux, C. Lorrie Epling, Elizabeth Sinclair, Maria A Suni, Kara Punt, Sandra Calarota, Sophia El-Bahi, Gailet Alter, Hazel Maila, Ellen Kuta, Josephine Cox, Clive Gray, Marcus Altfeld, Nolwenn Nougarede, Jean Boyer, Lynda Tussey, Timothy Tobery, Barry Bredt, Mario Roederer, Richard Koup, Vernon C Maino, Kent Weinhold, Giuseppe Pantaleo, Jill Gilmour, Helen Horton, and Rafick P Sekaly. Standardization of cytokine flow cytometry assays. *BMC Immunol*, 6:13, 2005. doi: 10.1186/1471-2172-6-13. URL `http://dx.doi.org/10.1186/1471-2172-6-13`.

D. Sarkar, N. Le Meur, and R. Gentleman. Using flowViz to visualize flow cytometry data. *Bioinformatics*, 24(6):878, 2008.